1983

On Improving Testing: A Student Evaluation Study

Marina Estabrook

Daniel L. Wick

On Improving Testing: A Student Evaluation Study

Marina Estabrook and Daniel L. Wick

Teaching Resources Center
University of California, Davis

Although testing is a crucial part of instruction, it is an area not easily accessible to instructional development centers for promoting test improvement.

The Teaching Resources Center, like other instructional development units, uses three major access routes to testing: (1) a computer test-scoring and test-analysis service, (2) workshops on testing, and (3) teaching consultation. But these approaches affect a relatively small number of faculty and address a limited aspect of testing, primarily the multiple choice exam.

As a step toward expanding our efforts in test improvement we sought to better acquaint ourselves and our faculty with our institution's testing practices and also with testing issues that are of concern to students. We therefore designed a study to provide an overview of the kinds of tests students take at the University of California at Davis (UCD), along with student evaluations of test formats. In addition, our study was designed to identify good and bad tests and testing practices from the perspective of students.

Method

A two-part questionnaire was sent to 200 randomly selected undergraduate students (see attached questionnaire). Part One asked students to describe and evaluate the tests they took the previous quarter. Because we estimated that students take 3-4 courses per

quarter, each student received four copies of Part One. By checking the appropriate box, students were asked to describe the course in terms of academic discipline, course level, class size, number of tests, including the final, and test format. In order to preserve the anonymity of the instructor, students were asked *not* to identify the instructor or the specific course. In addition to descriptions, students were asked to rate the exams using a five-point scale ranging from strongly agree to strongly disagree on the following evaluation items: (1) Exams accurately assessed what I learned in this course; (2) I learned a great deal from taking the exams in this course; (3) Exams were reasonable in difficulty; (4) I had sufficient time to complete the tests: (5) Exam questions were fair; (6) Exams covered a reasonable amount of material; (7) Exams stressed important points of the lecture/test; (8) Exam questions were free of ambiguity; (9) Grading practices in this course were fair; (10) The exams in this course were good overall; (11) My grade in this course accurately reflected my knowledge of the material.

Part Two asked students to report their best and worst testing experiences of their entire college career.

Results

Eighty-nine students (45%) responded with evaluation/description of tests in 305 courses. Respondents were equally distributed among freshmen (24), sophomores (20), juniors (25), and seniors (20). However, students rated approximately twice as many exams in lower-division (198) than upper-division courses (106).

Number of Exams Students Take

Overall, the median number of exams each student took per course, including the final, was 2.8 exams. Of the 305 courses described by students, no exams were administered in 10 of the courses (mostly studio classes).

The number of exams per course is related to class size, discipline, and exam format. Fewer exams are given in small than in large classes (Table 1), in Arts and Humanities than in other major disciplines

(Table 2), and in courses in which only essay tests are administered (Table 3).

TABLE 1
Median Number of Exams per Course and Class Size

Class Size	Median # of Tests	# of Courses
0 - 25	2.65	60
26 - 50	2.75	42
51 - 75	2.91	34
76 -100	3.20	26
101 -200	3.03	67
200 +	3.03	76
		305

$X^2 = 125.59$; df = 30; $p < .001$

TABLE 2
Median Number of Exams per Course and Discipline

Course Format	Median # of Exams	# of Courses
Arts & Humanities	1.84	79
Biological Sciences	3.03	49
Engineering	3.06	19
Physical Sciences	2.30	81
Social Sciences	2.67	67
Other	2.50	6
		301

$X^2 = 157.01$; df = 30; $p < .001$

TABLE 3
Median Number of Exams per Course and Exam Format

Course Format	Median # of Exams	# of Courses
Multiple Choice/True-False	3.08	26
Short Answer/Sentence Comp.	3.00	24
Essay	2.02	55
Problem Solving	3.15	65
Mixed Format	3.00	125
		305

$X^2 = 82.47$; df = 24; $p < .001$

Kind of Exams Students Take

The most prevalent test format to which students are exposed, regardless of class size, is the mixed format test that contains varying combinations of multiple choice, true-false, matching, sentence completion, short answer, essay, problem solving, and other formats. Of the 295 course exams evaluated, 42% were the mixed format type. Course exams that are strictly of one format are less prevalent: problem solving (22%), essay (19%), short answer (8%), and multiple choice (9%). Students took virtually no classes in which exams were all true-false, sentence completion, matching, or some other format unidentified on the questionnaire (Table 4).

TABLE 4
Frequency of Exam Formats

Type of Format	n Courses	% of Total
Multiple Chioce	24	9
True-False	2	0
Matching	0	0
Sentence Completion	1	0
Short Answer	23	8
Essay	55	19
Problem Solving	65	22
Mixed Format	125	42
Other	0	0
TOTAL=	295	100

To help establish what combination of test formats constitutes the mixed format test, test components were categorized in terms of response restrictiveness. Multiple choice, true-false and matching were classified as restricted response formats. Essay, short answer, sentence completion, and problem solving were classified as free response formats.

The results indicate that the most prevalent type of mixed format test is one that combines restricted and free response test items (62%). Although there is considerable variation in the mixture of these formats, the combination of short answer and multiple choice is the

most common. The least prevalent type of mixed format is one composed of only restricted response formats (3%) (Table 5).

TABLE 5
Types of Mixed Format Tests

Combination of Formats	n Courses	% of Total
Restricted and Restricted	4	3
Restricted and Free	76	62
Free and Free	43	35
TOTAL=	123	100

Evaluation of Course Exams Depending on Test Format

For purposes of analysis multiple choice and true-false exams were combined as were short answer and sentence completion. The results show that purely multiple choice/true-false exams, when compared to other test formats, consistently received the lowest median rating on all eleven evaluation items. The lower ratings were statistically significant on items pertaining to instructional value of the test, difficulty, and fairness of examination questions. However, the ratings of other test formats were fairly similar (Table 6). (See following page.)

Student dissatisfaction with multiple choice tests is further evidenced by student descriptions of best and worst testing experiences. Of the students that mentioned test format, 77% referred to multiple choice in connection with *worst* experiences but only 6% referred to multiple choice in connection with *best* experiences. Students describing best experiences mentioned essays (53%) and mixed format tests (32%) most frequently (Table 7).

Table 6
Median Rating of Course Exams by Test Format Based on a 5-point Scale Ranging from Strongly Agree (5) to Strongly Disagree (1)

	Multiple Choice/True-False (n=26)	Short/ Answer Sentence Completion (n=24)	Essay (n=55)	Problem Solving (n=65)	Mixed Format (n=125)	Significance*
1. Exams accurately assessed what I learned in this course.	3.30	4.25	4.2	3.73	3.8	.48
2. Exams in this course had instructional value.	2.30	4.05	3.92	3.26	3.27	.006[1]
3. Exams were reasonable in difficulty.	2.87	3.90	3.82	3.82	3.88	.01[2]
4. I had sufficient time to complete the test.	3.64	4.06	3.77	3.73	4.05	.08
5. Exam questions were fair.	2.50	3.94	3.71	3.75	3.60	.002[3]
6. Exams covered a reasonable amount of material.	3.88	4.16	3.91	3.96	4.05	.66
7. Exams stressed important points of the lecture/text.	3.35	4.12	4.02	4.06	3.78	.78
8. Exam questions were free of ambiguity.	2.50	3.50	3.89	3.55	3.39	.29
9. Grading prectices in this course were fair.	3.05	4.22	4.04	3.90	3.89	.12
10. The exams in this course were good overall.	2.70	4.00	4.06	3.81	3.63	.17
11. My grade in this course accurately reflected my knowledge of the material	2.83	4.14	3.86	3.75	3.61	.30

* Significance determined by Chi Square based on response pattern to each item.
1/ x^2 = 33.44; df = 16
2/ x^2 = 30.92; df = 16
3/ x^2 = 37.13; df = 16

TABLE 7
Student References to Test Format in Descriptions of Best and Worst Testing Experiences

Test Format	Best n Respondants	%	Worst n Respondants	%
Essay	28	53	6	14
Mixed	17	32	3	7
Multiple Choice	3	6	33	77
Problem Solving	5	9	1	2
TOTAL=	53	100	43	100

$x_2 = 51.28$; df. = 3; $p < .001$

Students' Best and Worst Testing Experiences

There was far more similarity among descriptions of best testing experiences than worst. When reporting the worst, students cited more personal experiences such as a calculator breaking down, acute test anxiety, coming late to the exam, etc., and this undoubtedly contributed to the variance. However, the results also suggest that there are more ways for tests to be bad than to be good.

Best Test Experiences

Among the best experiences the following themes recur and are listed here in order of prevalence. The best exams are those that (1) provide students with the optimal opportunity to demonstrate what they know and understand (n=18), (2) test students on understanding rather than on rote memorization of isolated facts (n=17), (3) address central points of lecture and reading (n=17), (4) ask straight-forward questions (n=12), (5) ask students to integrate the material they have learned (n=12), (6) cover a wide range of material (n=12), and (7) test students on a variety of skills (n=12). It is in connection with the above points that students most frequently mention a best exam that was either an essay or a mixed format test.

Additional characteristics to best testing practices cited by students are: (8) sufficient time to think about the test question, to complete the test, and recheck the answers (n=7), (9) instructors informing students in some fashion prior to the exam what they will

be expected to know, the kinds of questions they might be asked (analyze, apply, label, etc.), and what test format will be used. (Students find study questions particularly helpful for studying, promoting understanding, and learning much material thoroughly.) (n=7); (10) having the feeling that good performance on the test depends on how much students studied for the test and how well they knew the material (n=6).

Worst Test Experiences

Among the worst test experiences students identify the following, listed in order frequency:

1. ***Tricky questions.*** Students feel that some instructors design exams specifically to mislead students in order to obtain a wide bell-shaped curve. Such questions, students believe, are designed to measure test-wiseness and *not* knowledge and understanding. A number of students suggested that instructors make test questions progressively more difficult in order to achieve the designed wide distribution in test results.

2. ***Picky questions.*** Most students did not voice objection to being tested on detail per se, but rather on details that are unrelated to the major emphasis of the course. Students mention that so much material is covered in lecture and in the reading that it is impossible for them to memorize it all (n=18).

3. ***Confusing multiple choice items.*** Some major sources of confusion are: (a) selection of the best answer out of several more or less correct answers; (b) use of double negatives; (c) response formats such as a+b, a+b+c, all the above, or none of the above; (d) ambiguous questions and response choices; (e) errors on the test such as misspellings, omitting the correct answer, several correct answers, etc. (n=17).

4. ***Insufficient time.*** Students feel that some exams are too long and cover too much in too short a time. Furthermore, instructors often ask thought-provoking questions but give students no time to think (n=15).

5. ***Multiple choice exams as tests in problem solving.*** Students feel that multiple choice exams cannot fairly assess students' ability to solve problems nor can they provide partial credit (n=9).

6. ***Unfavorable classroom conditions.*** When taking the exam, students found it difficult to concentrate when the classroom was noisy, overcrowded, hot or stuffy (n=6).

7. ***Unwillingness of instructors to discuss exam results.*** Some instructors provide students with no opportunity to defend their answers. This criticism was cited primarily in connection with ambiguous multiple choice or true-false tests (n=2).

8. ***Instructors asking a broad essay question but looking for a specific answer.*** (n=2).

9. ***Instructors' unwillingness to give students any information prior to the exam about test format and what the test will cover.*** (n=2).

Conclusion

The results show that students took two mid-terms and a final in most courses. Contrary to our expectations, students took fewer exams in small than in large classes. This finding, however, is readily explainable in light of the fact that more essays are administered in small classes than in large, and that instructors give fewer exams when they administer essay exams than other test formats.

Our results also indicate that the most prevalent test format, regardless of class size, is the mixed format test composed of restricted and free response test items. Exams composed of only restricted response items, such as the multiple choice, true-false, and matching, are least prevalent.

With respect to student satisfaction with tests, the results indicate that mixed format, essay, short answer and problem solving tests are rated comparably. However, multiple choice exams as a sole measure of achievement were consistently rated the lowest by students, especially on items related to fairness of questions, difficulty, and instructional value.

Undoubtedly, multiple choice exams are among the most difficult to construct. They are also highly susceptible to criticism because the instructor is responsible for both the question and the answer. Student comments on multiple choice exams reflect some of the inherent difficulties associated with their construction: ambiguity, confusing wording, tricky and picky questions unrelated to the main emphasis

of the course, emphasis on factual information over conceptual understanding, etc. However, students do not only criticize test construction, but also question the appropriateness of the multiple choice test format for assessing certain cognitive skills, particularly problem solving.

Student descriptions of best and worst testing experiences,along with their ratings of test formats, suggest that students prefer methods of assessment that provide them with the optimal opportunity to demonstrate to themselves and to the instructor not only what they know but also what they understand. This is apparently best achieved by either a mixed format or a free response format test such as essay or short answer.

Implications for Instructional Development Centers

The testing survey has helped the Teaching Resources Center identify three important issues concerning testing improvement at UCD.

1. Contrary to our expectations, the pure multiple choice test was the test format least liked by students.

This implies that our Center should not place primary emphasis on the construction of multiple choice exams, but should assist faculty in developing such testing alternatives as the mixed format, essay, short answer and problem solving.

2. Although students were least satisfied with the pure multiple choice format, they regarded it more favorably when combined with other test formats, especially a free response format. This implies that our Center should encourage faculty members to use multiple choice exams only for testing straight-forward knowledge and to use free response formats for testing higher order learning. It is unrealistic to expect that faculty members will take the time and trouble to develop multiple choice items that adequately test such higher order learning skills as those identified in *Bloom's Taxonomy of Educational Objectives.* Student comments suggest that efforts by faculty to test such higher order skills by the multiple choice format often result in tricky, confusing, and ambiguous items.

3. Students were more satisfied with the essay exam format than we expected. Moreover, it was especially surprising to us that students did not complain about the subjective nature of the grading of essay exams. This suggests that while our Center should encourage the use of essay examinations, we need not expend special effort in assisting faculty in improving their grading of essays.

In conclusion, it should be pointed out that although the results of the testing survey at UCD may not be generalizable to other campuses, we strongly recommend that other instructional development centers undertake similar surveys to assist them in identifying the testing improvement issues relevant to their own institution.

TRC

STUDENT EVALUATION OF TESTING

Directions: The questionnaire is in two parts.

Part I asks you a general question concerning your UCD testing experiences.

Part II asks you specific questions concerning tests you took in each course in which you were enrolled *last quarter (Winter 1982)*. Four forms are included for this purpose; one for each course you took.

Part I - Entire UCD Experience

Without identifying a specific course or instructor, please *describe* your *best* and *worst* test experiences. Give as many examples in each category as you can. Refer to tests themselves, and test practices.

(a) *Best:*

(b) *Worst:*

Part II - Last Quarter Only

TRC

On the following pages please rate the tests in each course you took last quarter (Winter 1982). We assume you took approximately four courses.

Course #1 (do not identify course or instructor)

1. The course level was:
( ) lower-division ( ) upper-division ( ) graduate ( )other _____

2. The course falls into the following discipline category:
( )physical sciences ( ) engineering ( )biological sciences
( ) social sciences ( ) arts/humanities ( ) other _____

3. The class enrollment was approximately:
( ) 0-25 ( ) 26-50 ( ) 51-75 ( ) 76-100 ( ) 101-200
( ) 201 or more

4. Number of exams given in this course (including final) was:
( ) 0 ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 or more

5. The format of the exams was: (check more than one when applicable)
( ) multiple choice ( )essay ( )true-false ( )short answer
( ) sentence completion ( )matching ( )other _____

NA = not applicable	Strongly Agree		Neutral		Strongly Disagree	
	5	4	3	2	1	NA
6. Exams accurately assessed what I learned in this course	5	4	3	2	1	NA
7. I learned a great deal from taking the exams in this course	5	4	3	2	1	NA

NA = not applicable	Strongly Agree		Neutral		Strongly Disagree	
	5	4	3	2	1	NA
8. Exams were reasonable in difficulty	5	4	3	2	1	NA
9. I had sufficient time to complete the tests	5	4	3	2	1	NA
10. Exam questions were fair	5	4	3	2	1	NA
11. Exams covered a reasonable amount of material	5	4	3	2	1	NA
12. Exams stresed important points of the lecture/test	5	4	3	2	1	NA
13. Exam questions were free of ambiguity	5	4	3	2	1	NA
14. Grading practices in this course were fair	5	4	3	2	1	NA
15. The exams in this course were good overall	5	4	3	2	1	NA
16. My grade in this course accurately reflected my knowledge of the material	5	4	3	2	1	NA

17. *Comments:*